Instituto Complutense de Análisis Económico

A dominance approach for comparing the performance of VaR forecasting models

Laura Garcia-Jorcano
Department of Economic Analysis and Finance
(Area of Financial Economics)
Facultad de Ciencias Jurídicas y Sociales
Universidad de Castilla-La Mancha, Toledo, Spain

Alfonso Novales
Instituto Complutense de Análisis Económico (ICAE), and
Department of Economic Analysis, Facultad de Ciencias Económicas y Empresariales,
Universidad Complutense, 28223 Madrid, Spain

Abstract

We introduce three dominance criteria to compare the performance of alternative VaR forecasting models. The three criteria use the information provided by a battery of VaR validation tests based on the frequency and size of exceedances, offering the possibility of efficiently summarizing a large amount of statistical information. They do not require the use of any loss function defined on the difference between VaR forecasts and observed returns, and two of the criteria are not conditioned on any significance level for the VaR tests. We use them to explore the potential for 1-day ahead VaR forecasting of some recently proposed asymmetric probability distributions for return innovations, as well as to compare the APARCH and FGARCH volatility specifications with more standard alternatives. Using 19 assets of different nature, the three criteria lead to similar conclusions, suggesting that the unbounded Johnson SU, the skewed Student-t and the skewed Generalized-t distributions seem to produce the best VaR forecasts. The added flexibility of a free power parameter in the conditional volatility in the APARCH and FGARCH models leads to a better fit to return data, but it does not improve upon the VaR forecasts provided by GARCH and GJR-GARCH volatilities.

Keywords value at risk; backtesting; forecast evaluation; dominance; conditional volatility models; asymmetric distributions

JEL Classification C52; C58; G17; G32

Working Paper nº 1923
September, 2019

UNIVERSIDAD
COMPLUTENSE
MADRID

ISSN: **2341-2356**
WEB DE LA COLECCIÓN: http://www.ucm.es/fundamentos-analisis-economico2/documentos-de-trabajo-del-icae
Working papers are in draft form and are distributed for discussion. It may not be reproduced without permission of the author/s.

A dominance approach for comparing the performance of VaR forecasting models

Abstract

We introduce three dominance criteria to compare the performance of alternative VaR forecasting models. The three criteria use the information provided by a battery of VaR validation tests based on the frequency and size of exceedances, offering the possibility of efficiently summarizing a large amount of statistical information. They do not require the use of any loss function defined on the difference between VaR forecasts and observed returns, and two of the criteria are not conditioned on any significance level for the VaR tests. We use them to explore the potential for 1-day ahead VaR forecasting of some recently proposed asymmetric probability distributions for return innovations, as well as to compare the APARCH and FGARCH volatility specifications with more standard alternatives. Using 19 assets of different nature, the three criteria lead to similar conclusions, suggesting that the unbounded Johnson SU, the skewed Student-t and the skewed Generalized-t distributions seem to produce the best VaR forecasts. The added flexibility of a free power parameter in the conditional volatility in the APARCH and FGARCH models leads to a better fit to return data, but it does not improve upon the VaR forecasts provided by GARCH and GJR-GARCH volatilities.

Keywords: value at risk; backtesting; forecast evaluation; dominance; conditional volatility models; asymmetric distributions

JEL: C52; C58; G17; G32

1. Introduction

In spite of the recent emphasis placed on expected shortfall as an appropriate risk measure, the Basel committee emphasizes the evaluation of risk models through value at risk (VaR) backtesting. This is a sensible suggestion because of the difficulty in backtesting expected shortfall directly and also because expected shortfall estimates are calculated conditional on a VaR estimate.[1] The academic literature has long been interested in advancing sound proposals for the specification and estimation of financial risk through backtesting VaR models. To that end, a variety of statistical tests are usually applied to the alternative models considered, estimated for a number of assets, to select the best performing model over some out-of-sample period. In that exercise the researcher may compute a large number of test statistics, increasing with the number of model specifications, VaR tests, and assets considered. Summarizing such a large amount of information and using it to establish a preference ranking among models can easily become cumbersome.

To solve this problem, some authors [Braione and Scholtes (2016)] examine the percentage of tests in which each model has been rejected. Under such approach, all rejections receive the same weight with independence of the amount of sample evidence against the model. Other authors [Sarma et al. (2003)] follow a two-step strategy. In the first stage they select those models that pass some standard VaR tests based on the frequency of exceedances. In the second step, a loss function defined on the difference between observed returns and VaR forecasts is used to select the best among the models that passed the first stage. This two-step selection approach helps selecting a small subset among the competing models, but it could fail to identify some suitable models because they might have been removed in the first stage. Indeed, a model could be left out in the first stage because of failing to pass a given test at a specific confidence level, in spite of producing a smaller loss than other tests that have been judged to be statistically appropriate in the first stage. A further limitation with both model selection strategies is that the results are contingent on the significance level used for the tests, and we lack objective criteria for making that choice.

We follow a purely statistical approach to select the best VaR model, proposing three dominance criteria that use the information provided by VaR validation tests based on the frequency and size of exceedances. For any two models, the first criterion compares the tests in which each model has been rejected at a given significance level. The other two dominance criteria compare the extent to which the sample evidence is contrary to each model according to each VaR test. Rather than using a binary variable *reject/not reject* for each model, we rely on the amount of sample evidence against a given model under each test, as summarized in its *p*-value, to establish a ranking of VaR models.

[1] For recent work on expected shortfall backtesting see Novales and Garcia-Jorcano (2019), Du and Escanciano (2016), Acerbi and Szekely (2014), Righi and Ceretta (2013), among others.

We are essentially using the p-value as a measure of the distance between any two models in the probability space. This way, in the comparison between any two models, each test may receive a different weight in regard to the amount of sample evidence against each model. Contrary to most literature on risk model selection, two of the three dominance criteria that we propose do not depend on any significance level for the tests. We also show that it is indeed possible to summarize the vast amount of information that may emerge from the VaR tests in order to select a preferred VaR model.

Since the Fundamental Review of the Trading Book (FRTB) has recommended that expected shortfall be adopted as the main risk measure for the trading book under Basel III (Basel Committee on Banking Supervision, 2016), it is striking that backtesting is still largely based on VaR exceptions at the 99% level, for individual trading desks as well as for the whole trading book. Furthermore, the standard validation VaR tests are based only on the frequency of exceedances and their chronology (see, for example, Bhattacharyya and Ritolia, 2008; Corlu et al., 2016; Diamandis et al., 2011; Yu et al., 2010; Nozari et al., 2010; Bao et al., 2006; Mittnik and Paolella, 2000). A key limitation of such tests is that they do not distinguish between returns that are below but far from the VaR and those that are below but close to the VaR. A standard solution has consisted in considering some loss function defined on excess returns over and above the VaR (see, for example, Abad et al., 2016; ; Gerlach et al., 2011; Lee and Su, 2015; Louzis et al, 2014; Ozun et al., 2010). Various heuristic criteria have been proposed in the literature to support specific loss functions, usually penalizing models that produce exceedances with a large deviation from VaR (see, for example, Lopez, 1998, 1999; Sarma et al., 2003; Giacomini and Komunjer, 2005; Caporin, 2008). That approach may be appropriate when we want to select the model minimizing some economic cost, but it conveys only limited information for regulatory policy, since no normative rule can be deduced for the magnitude of these excess losses. Besides, the specification of the loss function will influence the selection of a best VaR model, so that except in situations when there are solid grounds for choosing a given functional form, the choice of loss function will be a component of model risk.

Loss functions and statistical tests have different goals and they should generally be expected to lead to a different selection of VaR models. Under the statistical approach that we follow, the target is to find the model that can best reproduce the tail of the return distribution, including the possibility of observing large exceedances with an associated low probability. To evaluate this aspect of a VaR model, new multivariate validation tests have appeared in literature that examine whether the size of VaR exceedances is as expected. They have been proposed as an implicit approach to backtesting Expected Shortfall estimates. Such tests are easy to understand, explain and implement, and they allow us to discriminate between models with different tail shape. In the proposed Dominance criteria we apply a model validation methodology based on these type of tests, along with more standard tests. By doing that, we follow the recommendation of the FRTB, which exhorts banks to go beyond the basic mandatory requirement to also consider more advanced backtests. We apply the Kupiec (1995) test, based on the number of VaR exceptions, the dynamic quantile test of Engle and Manganelli (2004) that examines their time dependence, and two tests that use the size of VaR exceedances, the Multinomial VaR backtest of Kratz et al., (2018), and the Risk Map test of Colletaz et al., (2013).

We consider three general volatility specifications with leverage, GJR-GARCH, APARCH, and FGARCH, as well as the standard symmetric GARCH model as benchmark. FGARCH allows for shifts and rotations in the news impact curve, and it includes as special cases the symmetric GARCH, GJR-GARCH, and APARCH models, thereby allowing for testing how simpler models fit the data. Furthermore, the FGARCH and APARCH models take the power on the conditional standard deviation of the innovations as a free parameter, allowing more flexibility in the response to the dynamics of volatility. As probability distributions for the innovations we compare the performance of the skewed Student-t distribution and the skewed generalized error distribution as introduced in Fernandez and Steel (1998), the unbounded Johnson S_U distribution, the skewed generalized-t distribution from Theodossiou (1998), and the generalized hyperbolic skew Student-t distribution from Aas and Haff (2006), using the normal and symmetric Student-t distributions as benchmarks. An interesting feature of our work is the consideration of 19 assets of different nature, including stock market indices, individual stocks, interest rates, commodity prices, and exchange rates. We calculate VaR forecasts following the parametric approach, with an AR(1) model being estimated for daily returns in all cases.

Our results show that asymmetric probability distributions for return innovations lead to improved VaR performance, while symmetric probability distributions are clearly inappropriate. For the wide array of financial assets considered, the unbounded Johnson distribution, the skewed generalized-t distribution and the skewed Student-t distributions dominate other asymmetric distributions. Improved VaR forecasts are also obtained when using volatility models that incorporate a leverage effect, with negative innovations having a larger impact on volatility than positive innovations of the same size. Even though APARCH and FGARCH volatility specifications better fit the whole distribution of return innovations, suggesting that the standard deviation, rather than the variance, should often be used to model volatility dynamics, VaR forecasts from GARCH and GJR-GARCH models are generally better. Our analysis is performed for 1-day ahead 1% VaR estimates. Results for other forecast horizons or for different probability levels for VaR may differ.

The remainder of the paper is organized as follows: in Section 2 we briefly describe the volatility models, the probability distributions and the VaR tests used in our analysis. In Section 3 we introduce the three Dominance criteria to rank VaR models. In Section 4 we present our data set. In Section 5 we report full sample estimates of the different models for stock market indexes and for individual stocks. In Section 6 we analyze the results obtained by application of the dominance approach and obtain the preferred VaR forecasting models. In Section 7 we report the results of the dominance approach by asset class. Finally, Section 8 concludes the paper.

2. Models and tests

Value at risk (VaR) is a simple risk indicator that measures what loss will be exceeded only a small percentage of times in the next h trading days $(100\alpha\%)$. We define VaR as a quantile of the profit/loss distribution for a given horizon and a given shortfall probability, reporting VaR as a negative number. Thus, given the log-return $r_{t,t+h}$ of an asset between t and $t+h$, VaR at a level α is defined by $Pr(r_{t,t+h} < VaR_{\alpha,t+h}) = \alpha$.

We examine the performance of alternative variance specifications and probability distributions for return innovations in 1-day ahead VaR forecasting. A detailed description of the volatility specifications, probability distributions and VaR tests used in the paper is presented in the Appendix.

Together with the normal (N) and Student-t (ST), we consider probability distributions that are not often considered in the literature on VaR performance, such as the skewed Student-t distribution (SKST) and the skewed generalized error distribution (SGED) [Fernandez and Steel (1998)], Johnson SU distribution (JSU) [Johnson (1949)], skewed generalized-t (SGT) [Theodossiou (1998)] and generalized hyperbolic skew Student-t distribution (GHST) [Aas and Haff (2016)]. In the VaR literature, Johnson distributions are suggested in Zangari (1996), Mina and Ulmer (1999), the RiskMetrics Technical Document (1996), and Choi and Nam (2008). Simonato (2011) documents the performance of the Johnson system for expected shortfall computation, relative to closely competing approaches such as the Gram-Charlier and Cornish-Fisher approximations. The GHST distribution has been employed very little in financial applications because its estimation is computationally demanding. Nakajima and Omori (2012) use it to perform a Bayesian analysis of a stochastic volatility model. Among multivariate applications, Hu (2005) develops a method to calibrate a multivariate generalized hyperbolic distribution using the EM algorithm. Paolella and Polak (2015) also use the generalized hyperbolic distribution in a multivariate time series context. Leccadito et al. (2014) have compared the performance of a variety of volatility specifications and asymmetric distributions using multilevel VaR tests that apply independence and conditional coverage tests at different confidence levels.

While the need to consider asymmetric probability distributions for return innovations seems to be well established at this point, the preference for a given volatility specification is less clear. We consider three general volatility models with leverage, GJR-GARCH, APARCH, and FGARCH, with a standard symmetric GARCH model as benchmark. The FGARCH, APARCH, and GJR-GARCH models include a leverage term, allowing for a negative shock to have a greater impact on volatility than a positive shock of the same size. The FGARCH volatility, introduced by Hentschel (1995), is a complex and flexible specification that subsumes some of the most popular GARCH models, like the symmetric GARCH, GJRGARCH, and APARCH. To the best of our knowledge, there are no papers examining the performance of this model for VaR forecasting. The APARCH and FGARCH specifications deal with the power in conditional standard deviation as a free parameter.

We consider four VaR tests that use the number of exceptions (the unconditional coverage test by Kupiec, 1995), their independence (the dynamic quantile test of Engle and Manganelli, 2004), and their size (Multinomial VaR backtest of Kratz et al., 2018, and the Risk Map test of Colletaz et al., 2013). Kupiec (1995) proposed a simple binomial test for the number of VaR exceptions that it is often described as a test of unconditional coverage. Engle and Manganelli (2004) propose a test that also explicitly examines the independence of exceptions. Kratz et al. (2018) develop a natural extension to standard VaR backtesting that allows for testing VaR estimates at N different probability levels, thereby providing an implicit backtest for Expected Shortfall. We select 3 levels ($\alpha_1 = 1\%$, $\alpha_2 = 0.5\%$ and $\alpha_3 = 0.25\%$). Among the tests proposed by Kratz, et al. (2018), we use the Nass test, which performs better with small cell counts. The test statistic follows a multinomial distribution. Colletaz et al. (2013) propose a method that jointly accounts for the number and the magnitude of extreme losses, providing a graphical summary of all the information about the performance of a risk model. It relies on the concept of a *super exception*, which is defined as an exception whose loss exceeds not only $VaR(\alpha)$, but also $VaR(\alpha')$, where α' is a probability level smaller than α (in our application we use $\alpha = 1\%$ for the standard VaR and $\alpha' = 0.5\%$ for this test). An abnormally high frequency of super exceptions would mean that the magnitude of the losses with respect to $VaR(\alpha)$ is too large. The approach consists on jointly testing for the number of VaR exceptions and superexceptions. A detailed description of the tests considered in the paper is presented in the Appendix.

3. Dominance among VaR models

Our first approach to dominance requires that the dominant model be rejected less often than the dominated model, but also a comparability condition that in a large proportion of cases when the dominant fails to pass the test, the dominated model also fails. We use M1 and M2 as a general notation to refer to any pair of models under comparison.[2]

Definition 1. *Let β be some specified threshold between 0 and 1, let R_1 and R_2 be the number of VaR validation tests in which the models M1 and M2 are rejected at a chosen significance level p, and let r_{12} be the number of common rejections for both models. We say that model M1 is dominated by*

[2]Along the paper we think of a VaR model as a combination of a probability distribution and a volatility specification for return innovations.

model M2 if i) M1 has been rejected in at least as many tests as M2, and ii) $q = r_{12}/R_2 \geq \beta$. If M2 dominates M1, the intensity of that dominance is measured by the ratio $I1_{12} = q \cdot R_1/R_2$.

Notice that β does not need to be related to the significance level p at which VaR validation tests are implemented. With $\beta = 1$ we would have a transitive relationship among VaR models, although the implied condition for dominance would then be too strong to be satisfied in practice. If $\beta = 0$, then M2 would dominate M1 whenever $R_2 < R_1$, which defines again a transitive relation but it would be too weak to be meaningful. If $\beta = 0$ almost any two models are comparable. With $\beta = 1$ it would be hard to find two models that are comparable. If the number of rejections of M1 is twice as large as that of M2, and all the rejections of M2 are also rejections of M1, then $I1_{12} = 2$. If half of the rejections of M2 were rejections of M1, then $I1_{12} = 1$, and if there were no common rejections, then $I1_{12} = 0$, and no model would dominate the other. The spirit of this criterion is that for dominance, the q-ratio should approach 1. In our analysis the ratio R_1/R_2 was usually below 5, except for a few exceptions, and it was generally between 1 and 2. Hence, the indicator $I1_{12}$ falls in most cases between 1 and 5, with values below 1 indicating weak dominance.

The other two dominance criteria are based on the amount of sample evidence against each model. In the four VaR validation tests the null hypothesis is H0: the VaR model is 'appropriate', in some sense that is specific to each test. Since the p-value is the probability that a sample with similar characteristics to ours would produce a test outcome more contrary to H0 than the one we have, it is natural to think of a model with a higher p-value in VaR validation tests as a model with a stronger support from the sample data. Under the two dominance criteria we introduce next, which are based on test p-values, model M2 could dominate model M1 even if $R_1 < R_2$.

Definition 2. *Let N_2 be the number of tests in which the p-value for M2 is higher than for M1 and let N_1 be the number of tests in which the opposite happens, with $I2_{12} = N_2 - N_1$. M2 dominates M1 if $I2_{12} > 0$, while M1 dominates M2 if $I2_{12} < 0$. The intensity of dominance is given by the absolute value of $I2_{12}$.*

The comparison between VaR models in this definition has the added value of not being contingent on any significance level, since we consider all the VaR tests that we have run. If desired, the definition could incorporate a significance level p and construct the $I2_{12}$ indicator using just the set of tests that lead to rejection of either model, M1 or M2, at that significance level. Using $p = 0$ would amount to considering all the tests in the construction of the indicator, as in the definition above.

The third approach takes into account the degree to which the sample evidence contraindicates each model, as measured by the p-value of each test. The definition could be extended to consider any monotone, bounded function of the differences in p-values.

Definition 3. *Let $pv1_i$ and $pv2_i$ be the p-values for models M1 and M2 in test i, $i = 1,2,...,N$, and denote by S_1 the sum of the differences in p-values in tests when $pv1_i > pv2_i$, and by S_2 the sum of differences in p-values when $pv1_i < pv2_i$, and compute the indicator $I3_{12} = S_2 - S_1$. Model M2 dominates M1 if $I3_{12} > 0$, and M1 dominates M2 if $I3_{12} < 0$. The intensity of dominance is given by the absolute value of $I3_{12}$.*

4. The data

We work with daily percentage returns for five groups of assets of different natures over the sample period 01/04/2000-12/31/2015 (4173 observations). Daily returns are computed as 100 times the first difference of log prices, i.e. $100[\ln(P_{t+1}) - \ln(P_t)]\%$. The financial assets considered are: the stock market indices IBEX 35 (€), NASDAQ 100 ($), FTSE 100 (£), and NIKKEI 225 (¥); the individual stocks IBM ($), SAN (€), AXA (€), and BP (£); the interest rate IRS 5Y (€), interest rate of GERMAN BOND 10Y (€), and interest rate of US BOND 10Y($); the commodity prices CRUDE OIL BRENT ($ per barrel), NATURAL GAS ($ per Million British Thermal Units), GOLD ($ per Troy Ounce), and SILVER (Cents $ per Troy Ounce); and the exchange rates EUR/USD (€), GBP/USD (£), JPY/USD (¥), and AUD/USD (Australian $). The data were extracted from Datastream.

Table 1 reports descriptive statistics for daily returns. All the assets have mean and median returns close to zero. Returns on interest rates are reported as the log changes in the price of implicit zero coupon bonds having the value of an interest rate as a yield. In terms of standard deviation, the sample range is higher for AUD/USD (18.7), IRS (18.0), and US BOND (17.1) and lower for JPY/USD (13.2), EUR/USD (13.4), SILVER (13.8), and the interest rate on the GERMAN BOND (13.9). The unconditional standard deviation is similar for assets in the same class, except for commodities, where GAS (4.19) and OIL BRENT (2.28) are more volatile than GOLD (1.13) and SILVER (1.93). NASDAQ is more volatile than other stock market indices and AXA is the most volatile stock. The $US exchange rate for the Australian dollar has a higher standard deviation than do those for the euro, British pound, or Yen. AUD/USD, SILVER, GOLD, and NIKKEI have significant negative skewness, while GAS, AXA, JPY/USD, and NASDAQ have high positive skewness. For all the assets considered the kurtosis is high, implying that the return distributions have tails much thicker than the normal distribution. The kurtosis is especially large for AUD/USD, GAS, IBM, and AXA, while the interest rate of the GERMAN BOND and the JPY/USD exchange rate have lower kurtosis. Together with a large sample size, these values for skewness and kurtosis yield a very large Jarque-Bera statistic, leading to the rejection of the assumption of normality in all cases.

5. Parameter estimates

To perform a VaR analysis we estimate the four volatility models GARCH, GJR-GARCH, APARCH, and FGARCH under each of the different probability distributions assumed for the innovations: normal, Student-t, skewed Student-t, skewed generalized error, unbounded Johnson S_U, skewed generalized-t, and generalized hyperbolic skew Student-t. To save space, we only report in Table 2 full sample estimates for the JSU-FGARCH model for stock market indices and individual stocks.[3] The results for the other models and asset classes are available from the authors upon request. An AR(1) model was specified for the conditional mean return in all cases. Most computations were performed with the R (version 3.1.1) package rugarch (version 1.3-4), designed for the estimation and forecast of various univariate ARCH-type models. The exception is the estimation of models under the skewed generalized-t and the generalized hyperbolic skew Student-t distributions, for which we used the sgt package (version 2.0) and the SkewHyperbolic package (version 0.3-2), respectively.

The parameter estimates show the FGARCH model to be successful in capturing the autocorrelation and heteroscedasticity exhibited by the data. The Ljung-Box Q statistic for five lags computed on the standardized residuals does not show evidence of autocorrelation at the 1% significance level. The same statistic computed with nine lags on the squared standardized residuals is not significant at 1%. The autoregressive effect in volatility is strong, with a β_1-parameter generally above 0.90, suggesting strong memory effects. The leverage effect is defined by η_{11} and η_{21}. The η_{11} incorporates rotation in the news impact curve. By allowing slopes of different magnitudes on either side of the origin, the news impact curve produces asymmetric variance response. A positive value of η_{11} corresponds to a clockwise rotation and implies that negative shocks increase volatility more than positive ones, as in IBEX, IBM, SAN, and BP. The η_{21} incorporates shifts in the news impact curve. A positive value of η_{21} causes a rightward shift of the news impact curve implying the type of asymmetry that matches the stylized facts of return volatility: negative shocks increasing volatility more than positive shocks of the same size.

Estimates of the skewness parameter (γ) of the Johnson S_U distribution are less than 0 for the market indices and individual stocks, suggesting the convenience of incorporating negative asymmetric features in the probability distribution in order to model innovations appropriately. Estimates of kurtosis parameters (δ) for the Johnson S_U distribution are greater than 2, suggesting the convenience of incorporating heavy tails in the probability distribution of the standardized innovations. Finally, we should emphasize that the λ-parameter for the FGARCH and the δ-parameter for APARCH take values between 0.97 and 1.54, differing significantly from 2 in most cases. This is in line with the

[3]The estimation results for APARCH volatility model under the different probability distributions for the stock market indexes and for individual stocks in our sample are reported in Tables A1 and A2 of the Online Appendix.

results of Taylor (1986), Schwert (1990), and Ding et al. (1993) that indicate that there is substantially more correlation among absolute returns than among squared returns, a reflection of the 'long memory' of high-frequency financial returns. Our estimates of the FGARCH and APARCH models for the different asset classes (shown in the online Apppendix) suggest that, contrary to standard practice, we should model the conditional standard deviation for stock market indices, individual stocks, and metals, the conditional variance ($\lambda = \delta = 2$) for interest rates, and a value between conditional standard deviation and variance ($\lambda = \delta = 1.5$) for energy commodities and exchange rates. In summary, these results indicate the need for a model featuring a negative leverage effect in the equation for conditional volatility combined with an asymmetric distribution for the underlying error term when representing stock market data. Furthermore, to get a good fit to sample return data that equation probably should be specified for the conditional standard deviation, rather than for the conditional variance. However, we will see below that such a better fit does not lead to better VaR forecasts.

In the lower panel we report the log-likelihood values of the four volatility models (GARCH, GJR-GARCH, APARCH, and FGARCH) under the different probability distributions. For individual stocks as well as for stock indices the least restricted FGARCH model achieves the highest likelihood, followed by the APARCH, GJR-GARCH, and GARCH models, as was expected. Furthermore, the differences between them are statistically significant in most cases. The results for all the other assets are similar. Over the whole sample of assets, the likelihood difference between the APARCH and FGARCH models is often not significantly different from zero. The GARCH specification was rejected when compared with the GJR-GARCH, and the latter rejected against the APARCH specification in 15 out of the 19 assets. Another observation is that for the two 10-year interest rates, none of the volatility specifications is rejected in a pairwise comparison under a likelihood ratio test. It was also difficult to discriminate among volatility specifications for exchange rates.[4]

6. Selecting the best VaR models

We analyze in this section the performance of our estimated models to forecast the 1-day ahead, 1% VaR. The $\alpha = 1\%$ significance level for VaR is a compromise between trying to capture extreme events and avoiding an excessively low number of exceptions, but results may change for different significance levels as well as for longer forecast horizons. Considering the left tail is not a trivial choice either, since results for both tails may differ significantly for asymmetric return distributions. We estimate the one-period VaR parametrically as $VaR_{\alpha,t} = \mu_t(\theta) + \sigma_t(\theta)F^{-1}(\alpha|\theta)$, where $\mu_t(\theta)$ represents the conditional mean, $\sigma_t(\theta)$ is the conditional standard deviation, and $F^{-1}(\alpha|\theta)$ denotes the corresponding quantile of the distribution of the standardized innovations z_t at a given $\alpha\%$ signif-

[4]With the only exception of the Australian dollar.

icance. After that, we evaluate the performance of VaR models using the four tests mentioned in the previous section.

Our goal is to evaluate the adequacy of the different VaR forecasting models by the amount of sample evidence against each model in each test, as indicated by its p-value. The problem is that running even a few tests for a wide array of models and a variety of assets leads to a vast amount of information from which it is not easy to derive useful inferences. In our case, we have four tests to apply to each of the 28 combinations of a probability distribution for return innovations and a specification for the dynamics of their volatility. Each model is estimated for the 19 assets, so we end up with a total of 2,128 test statistics. We need to summarize this information in order to be able to draw some conclusion on the relative merits of the different probability distributions and volatility specifications as part of a valid value at risk model.

Tables A3-A6 in the Online Appendix show for each asset the number of observed violations of VaR forecasts, the statistic and p-value of each test for each combination of volatility model and probability distribution for the innovations. In all cases we compute out-of-sample VaR forecasts over the last five years in the sample: 2011-2015 (1260 data observations). Every day we receive new return data and compute 1-day ahead 1% VaR, estimating each model every 50 days. This frequency of estimation aims at reducing the computational cost as well as avoiding frequent parameter variation that might be due in part to just sample noise.

The three dominance criteria are based on pairwise comparisons that would produce a 28x28 table displaying the value of the indicator for each pair of models (M1,M2). Glancing over such a large table will not produce conclusive results, so we summarize the information as follows: For a given Dominance criterion, we start by defining a subset of *dominant* models. The *preferred* VaR models will be those that perform better among the set of dominant models. Considering the performance among the dominant models helps to discriminate better which models should be preferred.

The first Dominance criterion is the only one for which we need to choose a significance level for the VaR tests. Applying the VaR tests at the 1% significance level, the first three columns in Table 3 display the number of pairwise comparisons in which each model is dominant, for different values of the β-ratio. As expected from the definition of this first criterion the number of dominance relationships increases as we decrease the value of β. Five models[5] consistently appear as being often dominant in pairwise comparisons for the different values of β, which is reassuring. Furthermore, these five models are very rarely dominated by other models, as shown in columns 4-6 of the Table for the three values of the β ratio.[6] To be safe, we choose $\beta = 0.75$ and consider as dominant all the

[5] SKST-FGARCH, JSU-GJRGARCH, JSU-APARCH, JSU-FGARCH, and SGT-APARCH

[6] Other model specifications are dominated in just a few cases, like SKST-GARCH, JSU-GARCH, SGT-GARCH, and SGT-GJRGARCH, but they do not come up as dominant models in the first three columns.

models that have a net positive balance according to this criterion. In a second step we consider pairwise comparisons just among the dominant models, and select those models with a positive average value of the Dominance indicator ($I1_{ij}$) across this reduced group of models, as indicated in boldface in column 8 of the Table. These are the four models with the SKST and JSU probability distributions, plus the SGED-APARCH, SGT-GARCH, SGT-GJRGARCH and SGT-APARCH models.

The frequency of test rejections is lower when we run tests at the 5% significance level (not shown in the paper) and so does the number of dominance relationships, unless we use a low value of the β ratio. In any event, the choice of the most preferred models would not differ much from the one made when VaR tests are run at 1% significance.

The statistic used in Definition 2 measures the number of tests that the p-value for model M1 is higher than for M2, minus the number of tests when the opposite happens. The idea is that a positive statistic indicates that the sample evidence is less contrary to M1 than to M2, while a negative statistic would indicate the opposite situation. Hence, for each of the 28 models we have a set of 27 $I2_{12}$ indicators. The first column in Table 4 displays for each model the median value of the set of 27 $I2_{12}$ indicators, and we consider as dominant the 13 models with a positive median for $I2_{12}$ (indicated with a boldface figure in column 1).[7] Among them, we select those models with a positive average value of the $I2_{ij}$ indicator in pairwise comparisons between dominant models, such an average being shown in the second column of the table. The results suggest a clear preference for the JSU-GARCH model, followed by the SKST-GARCH, JSU-GJRGARCH, JSU-FGARCH and SGT-GJRGARCH models. The dominance relationships between these 13 models are summarized in Figure 1. The best models are those where many arrow heads end up, and more so if the numbers on those arrows are high. Furthermore, we can see in the last two columns of the table that these models are the ones that are dominant more often and are dominated less often according to this criterion.

Table 5 shows the information needed to establish dominance comparisons according to Definition 3. We now have 14 models with a positive value for the median of the 27 indicators $I3_{12}$ (column 1 in Table 5). Twelve of these dominant models were also dominant models under Definition 2, so the consistency between the results of both definitions is quite high, which is again reassuring. Average values of the indicator across the dominant models, shown in column 2, leads again to JSU-GARCH as clearly being the preferred model. SKST-GARCH, SKST-FGARCH, SGED-GJRGARCH, JSU-GJRGARCH and SGT-GJRGARCH show also a good performance against the dominant models. As shown in columns 3 and 4 these are again generally the models that come up as dominant in pairwise comparisons according to this criterion, while rarely being dominated. Dominance between the 14 dominant models according to Definition 3 is summarized in Figure 2.

[7] Considering a positive mean indicator, rather than a positive median indicator to select the dominant models would not affect the choice of best VaR models.

As a summary of our dominance analysis, we can see that there is broad consensus from the application of the three Dominance criteria with respect to which VaR models should be preferred. Even when some specific choice needs to be done for the practical implementation of a given Dominance criterion, the selection of models is not too sensitive to the choice made. Table 6 summarizes the best performing models according to each Dominance criteria. The SKST-GARCH, JSU-GARCH, JSU-GJRGARCH and SGT-GJRGARCH models have been selected by each of the three dominance criteria, while models SKST-FGARCH and JSU-FGARCH have been selected by two of the Dominance criteria. Three of these six models incorporate the JSU probability distribution for the innovations, while the GARCH, GJRGARCH and FGARCH specifications for the conditional volatility of return innovations appear twice. It is interesting to also point out that the Normal, Student-t and Generalized Hyperbolic distributions were never part of a preferred model. Table 7.a shows the rank of each model under each dominance criterion, as well as the median and the average of those ranks. Table 7.b shows the average rank for each probability distribution for return innovations and each volatility specification. When using ranks we miss the information on the distance between the performance indicators of any two models but the six selected models also come on top as the highest ranked models according to the three criteria. The JSU distribution has clearly the lowest rank among the probability distributions[8] whereas GARCH and GJR-GARCH have the lowest rank among volatility specifications, now with much smaller differences than those among probability distributions.

7. Choosing a preferred VaR model by asset class

We have shown how the application of each of the three Dominance criteria leads to choosing a small set of models as the best ones for VaR estimation. In that process, we have used data for the 19 assets, so the chosen models are considered to be uniformly the best ones for the whole set of assets. In the risk management departments of financial institutions, where a large list of assets is evaluated frequently, it is clearly interesting to have a short list of models to be used for all assets. However, the preferred VaR models will depend on the characteristics of financial returns, and a given class of assets might have some specific characteristics that are not shared by other assets. Hence, it makes sense to search for the preferred VaR model for each asset class.[9]

Tables 8.a-8.e display the chosen models when applying each of the three Dominance criteria to each of the five asset classes. A difficulty now is that the set of chosen models is larger than when all the assets are used simultaneously, because the smaller number of tests that are run for each asset

[8] This is computed as the average of the median ranks of all models incorporating a given probability distribution for return innovations.

[9] Our test results could also be examined from the perspective of each asset, although the number of tests available is then considerably reduced and the ranking of models will be subject to significant sampling error.

class reduces the number of dominance relationships.[10] There are indeed some models selected for specific asset classes that do not show up when considering all the assets simultaneously. The N, ST, SGED, and GHST distribution appear now in the models chosen for some assets, and some of those models incorporate APARCH volatility. More importantly, Table 9.a shows the number of times that each model has been chosen as a preferred VaR model for a given asset class. The models chosen by the three criteria when considering the whole set of assets: SKST-GARCH, JSU-GARCH, JSU-GJRGARCH and SGT-GJRGARCH are also the models more often selected when considering each class of assets separately. Table 9.b pools together the choices made for the five asset classes by probability distribution and volatility specification. The JSU distribution is again the preferred one, while GARCH and GJRGARCH are the preferred volatility specifications.

8. Conclusions

This paper introduces a dominance approach to comparing a variety of VaR forecasting models. The three proposed dominance criteria use the information provided by a battery of VaR validation tests based on the frequency and size of exceedances. Two of the criteria are neither contingent on any significance level nor on the choice of any loss function, thereby avoiding some of the choices that need to be made when running VaR tests, which may end up conditioning the preference for one or other model. Using extensively the quantitative information provided by a large number of VaR tests and paying attention to the extent to which the sample evidence is contrary to each model, the use of either of these Dominance criteria can lead to robust conclusions on the preferred VaR models. Each of the three Dominance criteria establishes a rank among the alternative models purely in statistical terms, and it can be used by itself or combined with other criteria. In cases when there is solid ground to use a specific loss function defined on the size of VaR exceedances, the Dominance criteria could be used to select a small set of models whose VaR performance could then be compared on the basis of the chosen loss function.

We have used this approach to extend previous work on the forecasting performance of alternative VaR models by considering the volatility specifications GJR-GARCH, APARCH, and FGARCH and a set of asymmetric probability distributions, the skewed Student-t (SKST), skewed generalized error (SGED), unbounded Johnson (JSU), skewed generalized-t (SGT) and generalized hyperbolic skew Student-t distributions (GHST), several of them being relatively new in the financial literature. The normal and Student-t symmetric distributions and the GARCH volatility model are also used as benchmark. Our sample of daily data for 19 assets of different nature for the January 2000-December 2015 period covers the recent financial crisis of 2007-2009.

[10]In the same way than a shorter information set reduces the power of a statistical set.

We have applied four tests (the unconditional coverage test by Kupiec (1995), the dynamic quantile test by Engle and Manganelli (2004), the Multinomial VaR backtest of Kratz et al., (2018), and the Risk Map test of Colletaz et al., (2013) to the 28 models obtained when combining the 4 volatility models and the 7 probability distributions for return innovations. The number of test statistics quickly increases with the number of assets and models. With 19 assets, 4 tests and 28 models, we run 2,128 tests, but the Dominance criteria allow for an efficient summary of such a vast amount of information. Indeed, we end up with JSU-GARCH, JSU-GJRGARCH, SKST-GARCH and SGT-GJRGARCH as the preferred models for VaR estimation over the whole set of assets. Symmetric probability distributions are clearly inappropriate. Overall, JSU is the preferred probability distribution and GARCH and GJRGARCH are the preferred volatility specifications. In spite of offering a better in-sample fit, APARCH and FGARCH do not beat a more standard volatility specifications with leverage such as GJR-GARCH or even the simple GARCH model.

We have also searched for the best VaR model for each asset class, under the belief that assets in the same class share important statistical characteristics that may condition the performance of alternative VaR models. Indeed, we have found some differences in the specification of the preferred model for each asset class, but aggregating the choices made for each class, we end up again with the models chosen when the whole set of 19 assets was considered simultaneously. This should be seen as a desirable feature of our proposed procedure for selecting the best VaR models, although nothing in the way we have implemented our approach by asset class guaranteed the result.

The dominance criteria could be used for any set of tests that the researcher might consider interesting. The criteria could also incorporate a higher weight for those tests whose contrary evidence for a given model is considered to be more damaging for VaR forecasting. The second and third dominance criteria could also be extended to be based on the values of a given function of the p-values of each test for any two models, rather than using their raw values as we have done here. The criteria could also be used to choose among a set of forecasting models of any kind (macroeconomic, climatological), when forecasts from the different models are required to fulfill some testable conditions.

Bibliography

[1] Aas, K., and Haff, I.H. (2006). The generalized hyperbolic skew student'st-distribution. *Journal of Financial Econometrics*, 4(2), 275-309.

[2] Abad, P., Benito, S., Lopez, C., and Sanchez-Granero, M.A. (2016). Evaluating the performance of the skewed distributions to forecast value-at-risk in the global financial crisis. *Journal of Risk*, 18(5), 1-28.

[3] Abramowitz, M., and Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables (Vol. 9)*. Dover: New York.

[4] Acerbi, C. and Szekely, B. (2014). Backtesting Expected Shortfall. *Publication of MSCI*. https://www.msci.com/www/research-paper/research-insight-backtesting/0128184734.

[5] Bali, T.G., and Theodossiou, P. (2007). A conditional-SGT-VaR approach with alternative GARCH model. *Annals of Operations Research*, 151, 241-267.

[6] Basel Committee on Banking Supervision, Standards: Minimum Capital requirements for market risk (2016). *Bank for International Settlements*.

[7] Bao, Y., Lee, T., and Saltoglu, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *Journal of Forecasting*, 25, 101-128.

[8] Bhattacharyya, M., and Ritolia, G. (2008). Conditional VaR using EVT: towards a planned margin scheme. *International Review of Financial Analysis*, 17(2), 382-395.

[9] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307-327.

[10] Braione, M., and Scholtes, N.K. (2016). Forecasting Value-at-Risk under different distributional Assumptions. *Econometrics*, 4(3).

[11] Cai, Y., and Krishnamoorthy, K. (2006). Exact size and power properties of five tests for multinomial proportions. *Communications in Statistics-Simulation and Computation*, 35(1), 149-160.

[12] Caporin, M. (2008). Evaluating Value-at-Risk measures in the presence of long memory conditional volatility. *The Journal of Risk*, 10(3), 79-110.

[13] Choi, P., and Nam, K. (2008). Asymmetric and leptokurtic distribution for heteroscedastic asset returns: the SU-normal distribution. *Journal of Empirical finance*, 15(1), 41-63.

[14] Colletaz, G., Hurlin, C., and Pérignon, C. (2013). The Risk Map: A new tool for validating risk models. *Journal of Banking and Finance*, 37(10), 3843-3854.

[15] Corlu, C.G., Meterelliyoz, M., and Tiniç, M. (2016). Empirical distributions of daily equity index returns: A comparison. *Expert System with Applications*, 54, 170-192.

[16] Diamandis, P.F., Drakos, A.A., Kouretas, G.P., and Zarangas, L. (2011). Value-at-Risk for long and short trading positions: Evidence from developed and emerging equity markets. *International Review of Financial Analysis*, 20, 165-176.

[17] Ding, Z., Granger, C.W.J., and Engle, R.F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1, 83-106.

[18] Du, Z. and Escanciano, J.C. (2016). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 63(4), 940-958.

[19] Engle R.F., and Manganelli, S. (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, 22, 367-381.

[20] Fernandez, C., and Steel, M. (1998). On bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441), 359-371.

[21] Gerlach, R., Chen, C.W.S., Lin, E.M.H., and Lee, W.C.W. (2011). Bayesian forecasting for financial risk management, pre and post the global financial crisis. *Journal of Forecasting*, 31(8), 661-687.

[22] Giacomini, R., and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, 23(4), 416-431.

[23] Giot, P., and Laurent, S. (2003a). Value-at-Risk for long and short trading positions. *Journal of Applied Econometrics*, 18, 641-664.

[24] Glosten, L., Jagannathan R., and Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48, 1779-1801.

[25] Hentschel, L. (1995). All in the family nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, 39, 71-104.

[26] Hu, W. (2005). Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk. *Dissertation in the Florida State University*. http://diginole.lib.fsu.edu/islandora/object/fsu:181953/datastream/PDF/view

[27] Johnson, N.L. (1949). Systems of frequency curves generated by methods of translations. *Biometrika*, 36, 149-176.

[28] Kratz, M., Lok, Y.H., and McNeil, A.J. (2018). Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking and Finance*, 88, 393-407.

[29] Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 2, 174-184.

[30] Lambert, P., and Laurent, S. (2001). Modelling financial time series using GARCH-type models with a skewed student distribution for the innovations. *Mimeo, Université de Liege*.

[31] Leccadito, A., Boffelli, S., and Urga, G. (2014). Evaluating the accuracy of value-at-risk forecasts: New multilevel tests. *International Journal of Forecasting*, 30, 206-216.

[32] Lee, C.F., and Su, J.B. (2015). Value-at-Risk estimation via a semiparametric approach: Evidence from the stock markets. *Handbook of Financial Econometrics and Statistics*. Springer Science Business Media: New York.

[33] Lopez, J.A. (1998). Testing your risk tests. *Financial Survey* (May-Jun), 18-20.

[34] Lopez, J.A. (1999). Methods for evaluating Value-at-Risk estimates. *Federal Reserve Bank of San Francisco Economic Review*, 2, 3-17.

[35] Louzis, D.P., Xanthopoulos-Sisinis, S., and Refenes, A.P. (2014). Realized volatility models and alternative Value-at-Risk prediction strategies. *Economic Modelling*, 40, 101-116.

[36] McDonald, J. B., and Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric theory*, 4(3), 428-457.

[37] Mittnik, S., and Paolella, M. (2000). Conditional density and value-at-risk prediction of Asian currency exchange rates. *Journal of Forecasting*, 19, 313-333.

[38] Nakajima, J., and Omori, Y. (2012). Stochastic volatility model with leverage and asymmetrically heavy-tailed error using GH skew Student's t-distribution. *Computational Statistics and Data Analysis*, 56(11), 690-3704.

[39] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347-370.

[40] Nozari, M., Raei, S., Jahanguin, P., and Bahramgiri, M. (2010). A comparison of heavy-tailed estimates and filtered historical simulation: evidence from emerging markets. *International Review of Business Papers*, 6(4), 347-359.

[41] Ozun, A., Cifter, A., and Yilmazer, S. (2010). Filtered extreme-value theory for value-at-risk estimation: evidence from Turkey. *The Journal of Risk Finance*, 11(2), 164-179.

[42] Novales, A., and Garcia-Jorcano, L. (2019). Backtesting extreme value theory models of expected shortfall. *Quantitative Finance*, 19(5), 799-825. DOI: 10.1080/14697688.2018.1535182

[43] Paolella, M. S., and Polak, P. (2015). COMFORT: A common market factor non-Gaussian returns model. *Journal of Econometrics*, 187(2), 593-605.

[44] Righi, M.B. and Ceretta, P.S. (2013). Individual and flexible Expected Shortfall backtesting. *Journal of Risk Model Validation*, 7(3), 3-20.

[45] Riskmetrics, T. M. (1996). *JP Morgan Technical Document*.

[46] Sarma, M., Thomas, S., and Shah, A. (2003). Selection of value at risk models. *Journal of Forecasting*, 22, 337-358.

[47] Schwert, W. (1990). Stock volatility and the crash of '87. *Review of Financial Studies*, 3, 77-102.

[48] Simonato, J. G. (2011). The performance of Johnson distributions for computing value at risk and expected shortfall. *The Journal of Derivatives*, 19(1), 7-24.

[49] Taylor, S.J. (1986). Modelling Financial Time Series. *John Wiley and Sons, Inc*.

[50] Theodossiou, P. (1998). Financial data and skewed generalized t distribution. *Management Science*, 44, 1650-1661.

[51] Yu, P.L.H., Li, W.K., and Jin, S., (2010). On some models for value-at-risk. *Econometric Reviews*, 29(5-6), 622-641.

[52] Zangari, P. (1996). An improved methodology for measuring VaR. *RiskMetrics Monitor*, 2nd quarter, 7-25.

I. APPENDIX

I.1. Volatility models and probability distributions

Let x_t, for $t = 1,...,T$, be a time series of asset returns. It is convenient to break down the complete characterization of x_t into three components: (i) the conditional mean, μ_t; (ii) the conditional variance, σ_t^2; and (iii) the shape parameters, which determine the form of a conditional distribution (e.g. skewness or kurtosis) within a general family of distributions. Thus, we may write

$$x_t = \mu_t(\theta) + \varepsilon_t, \qquad \mu_t(\theta) = \mathbb{E}[x_t|\mathscr{F}_{t-1}] = \mu(\theta, \mathscr{F}_{t-1}), \qquad \varepsilon_t = \sigma_t(\theta) z_t,$$
$$\sigma_t^2(\theta) = \mathbb{E}[(x_t - \mu_t)^2|\mathscr{F}_{t-1}] = \sigma^2(\theta, \mathscr{F}_{t-1}), \qquad z_t \sim f(z_t|\theta).$$

The standardized innovation, $z_t = (x_t - \mu_t(\theta))/\sigma_t(\theta)$ has zero mean and unit variance. It follows a conditional distribution f with shape parameters that capture the possible asymmetry and fat-tailedness of returns, except in the case of the normal distribution. The vector θ contains all the parameters associated with the conditional mean and variance and the conditional distribution.

An AR(1) model for the conditional mean return is sufficient to produce serially uncorrelated innovations for all assets. For all the models we jointly estimate by the maximum likelihood method the parameters in the equation for the mean return, the equation for its conditional variance, and the probability distribution for the return innovations. The exception is the skewed generalized-t distribution, for which we use a two-step estimation method because of the numerical difficulty of estimating all its parameters jointly.[11]

I.1.1. Volatility models

The conditional variance of the GARCH(1,1) model [Bollerslev (1986)] is used as a benchmark, i.e.

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where $\omega > 0$, $\alpha_1, \beta_1 \geq 0$, $\alpha_1 + \beta_1 < 1$.

The standard GARCH model detects the existence of volatility clustering but it assumes that positive and negative error terms have the same effect on volatility. To incorporate asymmetric effects on volatility from positive and negative surprises, Glosten, Jagannathan, and Runkle (1993) proposed the GJR-GARCH(1,1) model, incorporating the negative impact of leverage in the conditional variance equation via the use of the indicator function $I(\varepsilon_{t-i} \leq 0)$, so that the variance equation becomes

$$\sigma_t^2 = \omega + \left[\alpha_1 \varepsilon_{t-1}^2 + \gamma_1 I(\varepsilon_{t-1} \leq 0) \varepsilon_{t-1}^2\right] + \beta_1 \sigma_{t-1}^2.$$

[11] In this case, we first estimated the AR(1)-GARCH conditional mean-volatility model assuming a generalized error distribution for the innovations, as suggested by Bali and Theodossiou (2007). The parameters of the skewed generalized-t distribution were estimated in a second stage using the standardized returns ($\frac{r_t - \phi_0 - \phi_1 r_{t-1}}{\sigma_t} = \frac{\varepsilon_t}{\sigma_t}$) obtained in the first step.

The volatility effect of a unit negative shock is $\alpha_i + \gamma_i$, while the effect of a unit positive shock is α_1. A positive value of γ_1 reflects that a negative innovation generates greater volatility than a positive innovation of equal size, and on the contrary for a negative value of γ_1.

The APARCH model (Asymmetric Power ARCH model) was proposed by Ding, Granger, and Engle (1993). This model incorporates volatility clustering, fat tails, excess kurtosis, the leverage effect, and the Taylor (1986) effect that the sample autocorrelation of absolute returns is usually larger than that of squared returns. The APARCH(1,1) variance equation is

$$\sigma_t^\delta = \omega + \alpha_1(|\varepsilon_{t-1}| - \gamma_1\varepsilon_{t-1})^\delta + \beta_1(\sigma_{t-1})^\delta,$$

where ω, α_1, γ_1, β_1, and δ are additional parameters to be estimated. The parameter γ_1 reflects the leverage effect $(-1 < \gamma_1 < 1)$. A positive (resp. negative) value of γ_1 means that past negative (resp. positive) shocks impact current conditional volatility more deeply than past positive (resp. negative) shocks. The parameter δ plays the role of a Box-Cox transformation of $\sigma_t(\delta > 0)$. The APARCH equation is supposed to satisfy the conditions: *i)* $\omega > 0$ (since the variance is positive), $\alpha_1 \geq 0$ and $\beta_1 \geq 0$, and, when $\alpha_1 = 0$ and $\beta_1 = 0$, then $\sigma_t^2 = \omega$; *ii)* $0 \leq \alpha_1 + \beta_1 \leq 1$. The APARCH model has great flexibility, having as special cases the GARCH and GJR-GARCH models, among others.

The FGARCH model of Hentschel (1995) is more general than the APARCH model, since it allows for the decomposition of the residuals in the conditional variance equation to be driven by different powers for z_t and σ_t. It also allows for both shifts and rotations in the news impact curve, where the shift is the main source of asymmetry for small shocks while rotation drives the asymmetry for large shocks. The FGARCH(1,1) is defined as

$$\sigma_t^\lambda = \omega + \alpha_1\sigma_{t-1}^\lambda f^\delta(z_{t-1}) + \beta_1(\sigma_{t-1})^\lambda,$$

where $f^\delta(z_{t-1}) = (|z_{t-1} - \eta_{21}| - \eta_{11}(z_{t-1} - \eta_{21}))^\delta$.

Positivity of $f^\delta(z_{t-1})$ is guaranteed when $|\eta_{11}| \leq 1$, which ensures that neither arm of the rotated absolute value function crosses the abscissa. However, the parameter η_{21} is unrestricted in size and sign. The magnitude and direction of a shift in the news impact curve are controlled by the parameter η_{21}, while the magnitude and direction of a rotation in the news impact curve are controlled by the parameter η_{11}. Other GARCH models permit only a shift or a rotation, but not both. By allowing for shifts in the news impact curve, the FGARCH model is more flexible than previous models, being able to capture asymmetries in volatility even in the presence of small shocks.

I.1.2. Probability distributions

As probability distributions for the innovations we compare the performance of the skewed Student-t, skewed generalized error (GED), unbounded Johnson S_U, skewed generalized-t (SGT), and generalized hyperbolic skew Student-t distributions, with the normal and symmetric Student-t distributions as benchmarks.

To account for the excess skewness and kurtosis typical of financial data, the parametric volatility models presented in the previous section can be combined with skewed and leptokurtic distributions for return innovations. The skewed Student-t distribution of Fernandez and Steel (1998) and Lambert and Laurent (2001)[12] is

$$f(z|\xi,\nu) = \frac{2}{\xi+\frac{1}{\xi}} s\{g[\xi(sz+m)|\nu]I_{(-\infty,0)}(z+m/s) + g[(sz+m)/\xi|\nu]I_{[0,\infty)}(z+m/s)\}, \qquad (1)$$

where $g(\cdot|\nu)$ is the symmetric (unit variance) Student-t density and ξ is the skewness parameter;[13] m and s^2 are, respectively the mean and the variance of the non-standardized skewed Student-t and are defined by

$$E(\varepsilon|\xi) = M_1(\xi-\xi^{-1}) \equiv m,$$

$$V(\varepsilon|\xi) = (M_2-M_1^2)(\xi^2+\xi^{-2}) + 2M_1^2 - M_2 \equiv s^2,$$

where $M_r = 2\int_0^\infty s^r g(s)ds$ is the absolute moment generating function. When $\xi = 1$ and $\nu > 2$, we have the skewness and the kurtosis of the (standardized) Student-t distribution. When $\xi = 1$ and $\nu = +\infty$, we get the skewness and kurtosis of the Gaussian density.

An alternative distribution for return innovations z_t that can capture skewness and kurtosis is the standardized skewed generalized error distribution, $SGED(0,1,\xi,\kappa)$, of Lambert and Laurent, with density[14]

$$f(z|\xi,\kappa) = \frac{2}{\xi+\frac{1}{\xi}} s\{g[\xi(sz+m)|\kappa]I_{(-\infty,0)}(z+m/s) + g[(sz+m)/\xi|\kappa]I_{[0,\infty)}(z+m/s)\},$$

where $g(\cdot|\kappa)$ is the symmetric (unit variance) generalized error distribution, ξ is the skewness param-

[12] Lambert and Laurent (2001) and Giot and Laurent (2003a) have shown that for various financial daily returns, it is realistic to assume that standardized innovations $\hat{z}_t$ follow a skewed Student-t distribution.

[13] The skewness parameter $\xi > 0$ is defined such that the ratio of probability masses above and below the mean is

$$\frac{Prob(z \geq 0|\xi)}{Prob(z < 0|\xi)} = \xi^2$$

[14] which is an extension of the generalized error distribution (GED) studied by Nelson (1991).

eter, κ represents the shape parameter, and $\Gamma(\cdot)$ is the gamma function. The mean (m) and standard deviation (s) are calculated in the same way as in the case of the skewed Student-t distribution. As κ increases the density gets flatter and flatter while in the limit, as $\kappa \to \infty$, the distribution tends toward the uniform distribution. Special cases are the normal distribution, when $\kappa = 2$, and the Laplace distribution, when $\kappa = 1$. For $\kappa > 2$ the distribution is platykurtic and for $\kappa < 2$ it is leptokurtic.

Another alternative is the Johnson S_U distribution. It was one of the distributions derived by Johnson (1949) based on translating the normal distribution by certain functions. Letting $Y \sim N(0,1)$, the standard normal distribution, the random variable Z has the Johnson system of frequency curves if it is a transformation of Y of the form $Y = \gamma + \delta g((Z-\xi)/\lambda)$. The form of the resulting distribution depends on the choice of function g. When $g(u) = sinh^{-1}(u)$, the resulting unbounded distribution is called the Johnson S_U distribution. The parameters of the distribution are ξ, $\lambda > 0$, γ, and $\delta > 0$.

We use a parameterization[15] of the original Johnson S_U distribution such that the parameters ξ and λ are the mean and standard deviation of the distribution. The parameter γ determines the skewness of the distribution with $\gamma > 0$ indicating positive skewness and $\gamma < 0$ negative skewness. The parameter δ determines the kurtosis of the distribution, and δ should be positive and most likely above 1.

The pdf of the Johnson's S_U, denoted here by $JSU(\xi, \lambda, \gamma, \delta)$, is defined by

$$f_Z(z) = \frac{\delta}{c\lambda} \frac{1}{\sqrt{(r^2+1)}} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}y^2\right],$$

with

$$y = -\gamma + \delta sinh^{-1}(r) = -\gamma + \delta \log\left[r + (r^2+1)^{1/2}\right],$$

$$r = \frac{z - (\xi + c\lambda\omega^{1/2} sinh\Omega)}{c\lambda},$$

$$c = \left\{\frac{1}{2}(\omega - 1)[\omega cosh 2\Omega + 1]\right\}^{-1/2},$$

where $\omega = exp(\delta^{-2})$ and $\Omega = -\gamma/\delta$. Note that $Y \sim N(0,1)$. Here $E(Z) = \xi$ and $Var(Z) = \lambda^2$.

The skewed generalized-t distribution proposed by Theodossiou (1998) is a quite flexible distribution extending the generalized-t distribution of McDonald and Newey (1988). It has probability density function

$$f(x|\mu,\sigma,\lambda,p,q) = \frac{p}{2v\sigma q^{1/p} B(\frac{1}{p},q)\left(\frac{|x-\mu+m|^p}{q(v\sigma)^p(\lambda sign(x-\mu+m)+1)^p} + 1\right)^{\frac{1}{p}+q}},$$

[15]This parameterization is used by the R package rugarch, which we use for estimating the parameters of our models.

with

$$m = \frac{2\nu\sigma\lambda q^{\frac{1}{p}} B\left(\frac{2}{p}, q - \frac{1}{p}\right)}{B\left(\frac{1}{p}, q\right)},$$

$$\nu = q^{-\frac{1}{p}} \left[(3\lambda^2 + 1) \left(\frac{B\left(\frac{3}{p}, q - \frac{2}{p}\right)}{B\left(\frac{1}{p}, q\right)} \right) - 4\lambda^2 \left(\frac{B\left(\frac{2}{p}, q - \frac{1}{p}\right)}{B\left(\frac{1}{p}, q\right)} \right)^2 \right]^{-\frac{1}{2}},$$

where $B(\cdot)$ is the beta function, and μ, σ, λ, p, and q are the location, scale, skewness, peakedness, and tail-thickness parameters, respectively, with $\sigma > 0$, $-1 < \lambda < 1$, $p > 0$, and $q > 0$. The skewness parameter λ controls the rate of descent of the density around $x = 0$. The parameters p and q control the height and tails of the density, respectively. The parameter q has the degrees of freedom interpretation in the case $\lambda = 0$ and $p = 2$.

The distributions belonging to the generalized hyperbolic family are more complex and novel. A special case of this family is the generalized hyperbolic skew Student-t distribution proposed by Aas and Haff (2006). This distribution has the important property that one tail has polynomial behavior while the other tail has exponential behavior. Further, it is the only subclass of the generalized hyperbolic family of distributions having this property. This is an alternative for modeling the empirical distribution of financial returns. It is often skewed, having one heavy and one semiheavy or Gaussian-like tail. The skew extensions to the Student-t distribution, like that of Fernandez and Steel, have two tails behaving polynomially. This means that they fit heavy-tailed data well, but they do not serve for modeling substantial skewness, since that requires one heavy tail and one non-heavy tail.

The probability density function of the generalized hyperbolic skew Student-t is given by

$$f_X(x) = \frac{2^{\frac{1-\nu}{2}} \delta^{\nu} |\beta|^{\frac{\nu+1}{2}} K_{\frac{\nu+1}{2}} \left(\sqrt{\beta^2(\delta^2 + (x-\mu)^2)} \right) exp(\beta(x-\mu))}{\Gamma(\frac{\nu}{2})\sqrt{\pi} \left(\sqrt{\delta^2 + (x-\mu)^2} \right)^{\frac{\nu+1}{2}}}, \qquad \beta \neq 0,$$

and

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\delta\Gamma(\frac{\nu}{2})} \left[1 + \frac{(x-\mu)^2}{\delta^2} \right]^{-(\nu+1)/2}, \qquad \beta = 0,$$

where $K_\nu(x) \sim \sqrt{\frac{\pi}{2x}} exp(-x)$ for $x \to \pm\infty$ is the modified Bessel function [Abramowitz and Stegun (1972)], and μ, δ, β, and ν determine the location, scale, skew, and shape parameters, respectively.

When $\beta = 0$ the density $f_X(x)$ is that of a noncentral Student-t distribution with ν degrees of freedom, expectation μ, and variance $\delta^2/(\nu-2)$.

I.2. VaR backtesting

The unconditional coverage test introduced by Kupiec (1995) is based on the number of violations, i.e. the number of times (T_1) that returns exceed the predicted VaR over a period of time T for a given significance level. If the VaR model is correctly specified, the failure rate ($\hat{\pi} = \frac{T_1}{T}$) should be equal to the prespecified VaR level (α). The null hypothesis $H0 : \pi = \alpha$ is evaluated through the likelihood ratio test

$$LR_{uc} = -2\ln\left(\frac{L(\Pi_\alpha)}{L(\widehat{\Pi})}\right) = -2\ln\left(\frac{(1-\alpha)^{T_0}\alpha^{T_1}}{(1-\hat{\pi})^{T_0}\hat{\pi}^{T_1}}\right) \quad \overset{T\to\infty}{\longrightarrow} \chi_1^2,$$

where $T_0 = T - T_1$.

The dynamic quantile test proposed by Engle and Manganelli (2004) overcomes some drawbacks of the conditional coverage test of (Christoffersen (1998)) using a linear regression model that links current violations to past violations. We define the auxiliary variable $Hit_t(\alpha) = I_t(\alpha) - \alpha$, so that $Hit_t(\alpha) = 1 - \alpha$ if $r_t < VaR_{t|t-1}(\alpha)$ and $Hit_t(\alpha) = -\alpha$ otherwise, where $I_t(\alpha)$ is equal to 1 if $r_t < VaR_{t|t-1}(\alpha)$ and equal to 0 otherwise. The null hypothesis of this test is that the sequence of hits (Hit_t) is uncorrelated with any variable that belongs to the information set Ω_{t-1} available when the VaR was calculated and it has a mean value of zero, which implies, in particular, that the hits are not autocorrelated. The dynamic quantile test is a Wald test of the null hypothesis that all slopes in the regression model

$$Hit_t(\alpha) = \delta_0 + \sum_{i=1}^{p} \delta_i Hit_{t-i} + \sum_{j=1}^{q} \delta_{p+j} X_j + \varepsilon_t,$$

are zero, where X_j are explanatory variables contained in Ω_{t-1}. The test statistic has an asymptotic χ^2_{p+q+1} distribution. In our implementation of the test, we use $p = 5$ and $q = 1$ (where $X_1 = VaR(\alpha)$) as proposed by Engle and Manganelli (2004). By doing so, we are testing whether the probability of an exception depends on the level of the VaR.

To account for both the number and magnitude of extreme losses, we also evaluate the performance of VaR models using the Multinomial VaR backtests (Kratz, et al., 2018) and the Risk Map test (Colletaz, et al., 2013).

The Multinomial VaR backtests introduced by Kratz, et al. (2018) are based on testing simultaneously VaR estimates at N levels leads to a multinomial distribution. Let $X_t = \sum_{j=1}^{N} I_{t,j}$, where $I_{t,j} = I_{r_t < VaR_{t|t-1}(\alpha_j)}$ is the exception indicator of the level α_j at time t, then the sequence $(X_t)_{t=1,\ldots,n}$ counting the number of VaR levels that are breached should satisfy two conditions: i) the unconditional coverage hypothesis, $P(X_t \leq j) = \alpha_{j+1}, \quad j = 0,\ldots,N \quad \forall t$, and ii) the independence hypothesis, X_t is independent of X_s for $s \neq t$. Let $MN(n,(p_0,\ldots,p_N))$ denotes the multinomial distribution with n trials, each of which may result in one of $N+1$ outcomes $0,1,\ldots,N$ according to probabilities $p_0,\ldots,p_N$ that sum to one. If we define observed cell counts by, $O_j = \sum_{t=1}^{n} I_{X_t = j}, j = 0,1,\ldots,N$, then, under the unconditional coverage and independence assumptions, the random vector $(O_0,...,O_N)$ should follow the multinomial distribution $(O_0,...,O_N) \sim MN(n,\alpha_1-\alpha_0,...,\alpha_{N+1}-\alpha_N)$, where $\alpha_0 = 0$ and $\alpha_{N+1} = 1$. More formally, let $0 = \theta_0 < \theta_1 < ... < \theta_N < \theta_{N+1} = 1$ be an arbitrary sequence of parameters and consider the model, $(O_0,...,O_N) \sim MN(n,\theta_1-\theta_0,...,\theta_{N+1}-\theta_N)$. We test null and alternative hypothesis given by,

$$
\begin{aligned}
&H_0 : \theta_j = \alpha_j \text{ for } \quad j = 1,...,N \\
&H_1 : \theta_j \neq \alpha_j \text{ for at least one } j \in \{1,...,N\}
\end{aligned}
$$

Various test statistics can be used to evaluate these hypothesis. Kratz, et al. (2018) propose three of five possible tests of multinomial proportions provided by Cai and Krishnamoorthy (2006): the standard Pearson chi-squared test, the Nass test, and the likelihood ratio test. We use the Nass test, which performs better with small cell counts, with a test statistic,

$$c \cdot S_N \overset{d}{\sim} \chi^2_\nu$$

with

$$S_N = \sum_{j=0}^{N} \frac{(O_{j+1} - n(\alpha_{j+1}-\alpha_j))^2}{n(\alpha_{j+1}-\alpha_j)}, \quad c = \frac{2\mathbb{E}(S_N)}{\mathbb{V}ar(S_N)} \quad \text{and} \quad \nu = c\mathbb{E}(S_N),$$

where $\mathbb{E}(S_N) = N$ and $\mathbb{V}ar(S_N) = 2N - \frac{N^2+4N+1}{n} + \frac{1}{n}\sum_{j=0}^{N} \frac{1}{\alpha_{j+1}-\alpha_j}$.

The Risk Map test introduced by Colletaz, et al. (2013) consists on the VaR backtest at two levels to account for both the number and magnitude of extreme losses. This approach exploits the concept of *super exception*, which they define as a loss greater than $VaR(\alpha')$, with α' smaller than α. An abnormally high frequency of super exceptions would suggest that the magnitude of the losses with respect to $VaR(\alpha)$ is too large. The approach consists on jointly testing the number of VaR exceptions and superexceptions,

$$H_0 : \mathbb{E}[I_t(\alpha)] = \alpha \quad \text{and} \quad \mathbb{E}[I_t(\alpha')] = \alpha'.$$

This joint null hypothesis can be tested using either a multivariate version of the unconditional coverage test (LR_{uc}). The authors follow Pérignon and Smith (2008) and define several indicator variables for revenues falling in each disjoint interval

$$J_{1,t} = I_t(\alpha) - I_t(\alpha') = \begin{cases} 1 & \text{if} \quad VaR_{t|t-1}(\alpha') < r_t < VaR_{t|t-1}(\alpha) \\ 0 & \text{otherwise} \end{cases}$$

$$J_{2,t} = I_t(\alpha') = \begin{cases} 1 & \text{if} \quad r_t < VaR_{t|t-1}(\alpha') \\ 0 & \text{otherwise} \end{cases}$$

and $J_{0,t} = 1 - J_{1,t} - J_{2,t} = 1 - I_t(\alpha)$. The $\{J_{i,t}\}_{i=0}^{2}$ are Bernoulli random variables equal to one with probability $1-\alpha$, $\alpha-\alpha'$, and α', respectively. However, they are clearly not independent since only one J variable may be equal to one at any point in time, $\sum_{i=0}^{2} J_{i,t} = 1$. We can test the joint hypothesis of the specification of the VaR model using a simple likelihood test. Let $N_{i,t} = \sum_{t=1}^{T} J_{i,t}$, for $i = 0,1,2$, the count variable associated with each of the Bernoulli variables. This multivariate unconditional coverage test is a likelihood ratio test LR_{MUC} that the empirical exception frequencies significantly deviate from the theoretical ones

$$\begin{aligned} LR_{MUC}(\alpha,\alpha') = & -2\ln\left[(1-\alpha)^{N_0}(\alpha-\alpha')^{N_1}(\alpha')^{N_2}\right] + \\ & + 2\ln\left[\left(\frac{N_0}{T}\right)^{N_0}\left(\frac{N_1}{T}\right)^{N_1}\left(\frac{N_2}{T}\right)^{N_2}\right] \xrightarrow{d} \chi_2^2 \end{aligned}$$

II. TABLES

II.1. STATISTICS

	Mean (bps.)	Median (bps.)	Max	Min	S.D.	Skewness	Kurtosis	J-B
IBEX	-0.47	2.89	13.48	-9.58	1.49	0.08	7.93	4234.84
NASDAQ	0.46	3.68	17.20	-11.11	1.85	0.19	9.62	7652.53
FTSE	-0.25	0	9.38	-9.26	1.21	-0.16	9.36	7042.80
NIKKEI	0.01	0	13.23	-12.11	1.50	-0.41	9.72	7979.58
IBM	0.42	0	12.26	-16.89	1.66	-0.07	11.63	12947.74
SAN	1.01	0	20.87	-15.19	2.19	0.15	9.11	6515.50
AXA	0.55	0	19.78	-20.35	2.67	0.27	10.09	8790.79
BP	-1.35	0	10.58	-14.04	1.71	-0.13	7.81	4041.28
IRS	0.55	0.48	1.92	-1.86	0.21	-0.28	8.53	5367.17
GER BOND	1.11	0.97	3.39	-2.33	0.41	-0.09	5.97	1536.83
US BOND	0.98	0.96	4.53	-5.57	0.59	-0.22	7.96	4307.77
BRENT	0.98	0	17.97	-18.72	2.28	-0.19	8.26	4831.81
GAS	0.01	0	37.81	-28.90	4.19	0.56	12.81	16946.14
GOLD	3.10	0.01	6.86	-10.16	1.13	-0.41	8.81	5991.49
SILVER	2.26	0	13.66	-12.98	1.93	-0.57	8.62	5724.23
EUR/USD	0.16	0	4.62	-3.84	0.63	0.14	5.48	1091.11
GBP/USD	-0.20	0	4.43	-3.88	0.57	-0.04	7.27	3170.80
JPY/USD	-0.41	-0.99	4.61	-3.71	0.63	0.27	6.96	2779.74
AUD/USD	0.23	1.86	6.70	-8.83	0.83	-0.82	15.13	26058.43

Table 1: Descriptive statistics for daily percentage returns. Mean and median returns are in basis points. SD denotes the standard deviation. J-B is the Jarque-Bera statistic to test for normality. Sample: 01/04/2000-12/31/2015.

II.2. PARAMETER ESTIMATES

JSU-FGARCH								
	IBEX	NASDAQ	FTSE	NIKKEI	IBM	SAN	AXA	BP
μ	0.00318	0.06282	-0.01138	0.00162	-0.00254	-0.01105	-0.01953	-0.02743
	(0.01572)	(0.01515)	(0.01247)	(0.01792)	(0.01754)	(0.02280)	(0.02736)	(0.02006)
ϕ_1	0.00606	-0.04216	-0.03840	-0.02369	-0.02572	-0.01661	0.02187	-0.01761
	(0.01538)	(0.01509)	(0.01572)	(0.015414)	(0.01425)	(0.01528)	(0.01539)	(0.01568)
ω	0.02206	0.01778	0.01695	0.04692	0.01655	0.03187	0.07130	0.02435
	(0.00357)	(0.00310)	(0.00272)	(0.00813)	(0.00532)	(0.00698)	(0.05009)	(0.00457)
α_1	0.03758	0.08337	0.07571	0.17079	0.07286	0.06935	0.06057	0.06743
	(0.01955)	(0.00495)	(0.02106)	(0.02071)	(0.01289)	(0.00667)	(0.00923)	(0.00880)
β_1	0.83352	0.76105	0.77693	0.81636	0.93089	0.90368	0.85899	0.91365
	(0.02517)	(0.01460)	(0.01833)	(0.01034)	(0.01610)	(0.01831)	(0.03528)	(0.00068)
η_{11}	0.05937	-0.66038	-0.24868	-0.40761	0.43700	0.45879	-0.04667	0.24458
	(0.15299)	(0.00013)	(0.07685)	(0.05517)	(0.01414)	(0.09180)	(0.14367)	(0.09374)
η_{21}	1.41500	3.42085	1.72560	1.31868	0.11635	0.39715	1.11610	0.46777
	(0.26515)	(0.04398)	(0.14478)	(0.07999)	(0.03628)	(0.05277)	(0.26664)	(0.05059)
$\lambda = \delta$	2.28353	1.83333	2.40002	1.69633	1.02395	1.35677	2.09148	1.30909
	(0.64471)	(0.30379)	(0.38457)	(0.22121)	(0.07678)	(0.24964)	(0.93286)	(0.18707)
γ skewness	-0.47165	-0.47566	-0.82536	-0.38057	-0.07064	-0.25156	-0.23318	-0.14377
	(0.11262)	(0.12768)	(0.19072)	(0.09789)	(0.05011)	(0.07957)	(0.10106)	(0.07341)
δ kurtosis	2.40499	2.62956	2.77958	2.17734	1.55715	2.08452	2.44055	2.03278
	(0.19263)	(0.25728)	(0.28522)	(0.16519)	(0.06743)	(0.13938)	(0.19061)	(0.12956)
Ljung-Box Test on Standardized Residuals Lag[5]								
statistic	2.1339	3.8390	3.0148	2.0270	1.8899	2.2545	5.1709	4.8835
p-value	0.6725	0.2522	0.4276	0.7036	0.7428	0.6373	0.0928	0.1165
Ljung-Box Test on Standardized Squared Residuals Lag[9]								
statistic	10.8310	11.8510	4.8760	3.0720	0.6810	11.4969	5.4916	1.7977
p-value	0.0331	0.0196	0.4478	0.7470	0.9958	0.0236	0.3609	0.9277
Log-Likelihoods								
FGARCH	-6810.914	-7137.549	-5686.932	-6970.236	-7032.462	-8289.825	-8907.877	-7473.873
APARCH	-6816.443	-7145.028	-5703.821	-6993.738	-7033.070	-8291.966	-8916.729	-7477.180
GJRGARCH	-6829.479	-7154.870	-5713.324	-7001.968	-7054.587	-8306.280	-8926.076	-7486.198
GARCH	-6899.894	-7213.013	-5796.144	-7031.025	-7068.898	-8352.110	-8965.944	-7509.518

Table 2: Parameter estimates of the FGARCH model for stock market indices and individual stocks under an unbounded Johnson probability distribution for return innovations. Estimated parameters are as indicated in the models shown in the Appendix. Standard deviations are reported in parentheses. The lower panel shows the log-likelihood values of the four volatility models considered in the paper.

II.3. DOMINANCE AMONG VaR MODELS

II.3.1. Definition 1 of Dominance

	Frequency of dominance ($p = 1\%$)							
	Number of times each model is						**Dominance**	**Average**
	dominant			**dominated**			**balance**	**indicator**
β	**0.9**	**0.75**	**0.5**	**0.9**	**0.75**	**0.5**	**for**	**among**
# of pairwise comparisons	*124*	*203*	*307*	*124*	*203*	*307*	$\beta = 0.75$	**Dominant models**
N-GARCH	1	2	2	21	22	25	-20	
N-GJRGARCH	2	3	3	12	20	23	-17	
N-APARCH	1	1	1	10	22	25	-21	
N-FGARCH	0	0	0	18	24	27	-24	
ST-GARCH	2	4	4	5	6	13	-2	
ST-GJRGARCH	3	6	8	**2**	9	17	-3	
ST-APARCH	5	5	6	5	11	18	-6	
ST-FGARCH	4	4	5	11	17	20	-13	
SKST-GARCH	4	9	13	**1**	**1**	5	**8**	**0.1**
SKST-GJRGARCH	7	13	17	**1**	4	5	**9**	**0.3**
SKST-APARCH	8	9	14	4	5	7	**4**	**0.2**
SKST-FGARCH	**13**	13	**25**	**0**	**0**	**0**	**13**	**1.2**
SGED-GARCH	3	5	12	4	5	10	0	
SGED-GJRGARCH	3	5	10	**3**	10	13	-5	
SGED-APARCH	6	11	14	4	4	5	**7**	**0.3**
SGED-FGARCH	3	7	9	6	6	8	**1**	0.0
JSU-GARCH	4	9	13	**1**	**1**	5	**8**	**0.1**
JSU-GJRGARCH	**13**	**18**	**23**	**0**	**0**	**1**	**18**	**0.7**
JSU-APARCH	**10**	**16**	**23**	**1**	**1**	**1**	**15**	**0.7**
JSU-FGARCH	**10**	**16**	**23**	**1**	**1**	**1**	**15**	**0.7**
SGT-GARCH	4	9	13	**1**	**1**	5	**8**	**0.1**
SGT-GJRGARCH	4	7	13	**0**	**1**	4	**6**	**0.1**
SGT-APARCH	**10**	**16**	**23**	**1**	**1**	**1**	**15**	**0.7**
SGT-FGARCH	3	7	9	6	6	8	**1**	0.0
GHST-GARCH	0	6	11	**0**	5	13	**1**	0.0
GHST-GJRGARCH	0	1	3	**0**	4	9	-3	
GHST-APARCH	0	0	4	6	13	22	-13	
GHST-FGARCH	1	1	6	**0**	**3**	16	-2	

Table 3: Dominance between $VaR_{1\%}$ models according to Definition 1. The table shows the frequency of dominance relationships from tests implemented at the 1% significance level. The left panel of the table displays the number of pairwise comparisons in which each model appears as dominant using 0.90, 0.75 and 0.50 as values for the β-ratio (boldface figures indicate models with at least 10, 15 and 20 dominance relationships). The right panel indicates the number of comparisons in which each model is dominated using 0.90, 0.75 and 0.50 as values for the β-ratio (boldface figures indicate models that are dominated by no more than other 3 models). Column 7 shows the net Dominance balance (the number of pairwise comparisons in which each model appears as dominant minus that one in which each model appears as dominated) for $\beta = 0.75$ (boldface figures indicate models that have a net positive balance). Column 8 shows the average value of the Dominance indicator, $I1_{ij}$ in pairwise comparisons among the dominant models. Boldface figures indicate models with a positive average indicator. These are the preferred models according to this Dominance criterion.

II.3.2. Definition 2 of Dominance

	All tests			
	Median	**Average**	**Dominant**	**Dominated**
N-GARCH	-56		1	26
N-GJRGARCH	-52		3	24
N-APARCH	-50		1	26
N-FGARCH	-46		1	26
ST-GARCH	-4		8	17
ST-GJRGARCH	-12		7	20
ST-APARCH	-16		4	22
ST-FGARCH	-17		4	22
SKST-GARCH	**21**	**8.2**	24	3
SKST-GJRGARCH	**9**	-3.2	20	7
SKST-APARCH	**4**	-9.3	15	12
SKST-FGARCH	-2		13	14
SGED-GARCH	**8**	-5.6	17	10
SGED-GJRGARCH	**6**	-3.4	20	6
SGED-APARCH	0		13	13
SGED-FGARCH	-6		11	15
JSU-GARCH	**26**	**16.7**	27	0
JSU-GJRGARCH	**19**	**7.0**	25	1
JSU-APARCH	**10**	-1.9	19	8
JSU-FGARCH	**18**	**4.5**	24	3
SGT-GARCH	**13**	-0.2	20	7
SGT-GJRGARCH	**14**	**2.2**	24	2
SGT-APARCH	**10**	-3.5	19	8
SGT-FGARCH	**6**	-11.5	14	12
GHST-GARCH	0		13	13
GHST-GJRGARCH	-13		7	19
GHST-APARCH	-13		7	20
GHST-FGARCH	-4		11	16

Table 4: Dominance between $VaR_{1\%}$ models according to Definition 2. Column 1 shows the median value of the set of 27 indicators $I2_{12}$ for each model. Column 2 shows the average indicator $I2_{12}$ using only pairwise comparisons between the dominant models. Column 3 shows the number of pairwise comparisons in which model is dominant. Column 4 shows the number of pairwise comparisons in which each model is dominated by other model. Boldface figures in the first two columns indicate models that have a positive value of the indicator. These are the preferred models according to this Dominance criterion.

II.3.3. Definition 3 of Dominance

	All tests			
	Median	**Average**	**Dominant**	**Dominated**
N-GARCH	-19.4		0	27
N-GJRGARCH	-16.7		1	26
N-APARCH	-16.5		2	25
N-FGARCH	-15		3	24
ST-GARCH	-1.6		10	17
ST-GJRGARCH	-3.1		8	19
ST-APARCH	-6.5		4	23
ST-FGARCH	-5.1		5	22
SKST-GARCH	**4.8**	**1.2**	26	1
SKST-GJRGARCH	**2.9**	-0.8	20	7
SKST-APARCH	**0.3**	-3.6	15	12
SKST-FGARCH	**4.2**	**0.6**	22	5
SGED-GARCH	**2.1**	-1.7	18	9
SGED-GJRGARCH	**4.4**	**0.8**	23	4
SGED-APARCH	-0.1		13	14
SGED-FGARCH	-1.9		9	18
JSU-GARCH	**15.1**	**12.3**	27	0
JSU-GJRGARCH	**4.7**	**1.1**	24	3
JSU-APARCH	**1.0**	-2.8	16	11
JSU-FGARCH	**3.2**	-0.5	21	6
SGT-GARCH	**2.4**	-1.3	19	8
SGT-GJRGARCH	**4.7**	**1.1**	25	2
SGT-APARCH	**1.4**	-2.4	17	10
SGT-FGARCH	-1.4		11	16
GHST-GARCH	**0.1**	-3.9	14	13
GHST-GJRGARCH	-3.9		6	21
GHST-APARCH	-3.4		7	20
GHST-FGARCH	-0.7		12	15

Table 5: Dominance between $VaR_{1\%}$ models according to Definition 3. Column 1 shows the median value of the set of 27 indicators $I2_{12}$ for each model. Column 2 shows the average indicator $I2_{12}$ using only pairwise comparisons between the dominant models. Column 3 shows the number of pairwise comparisons in which model is dominant. Column 4 shows the number of pairwise comparisons in which each model is dominated by other model. Boldface figures in the first two columns indicate models that have a positive value of the indicator. These are the preferred models according to this Dominance criterion.

II.3.4. Summary of Dominance Analysis

Panel 1

D1	D2	D3
SKST-GARCH	**SKST-GARCH**	**SKST-GARCH**
SKST-GJRGARCH		
SKST-APARCH		
SKST-FGARCH		**SKST-FGARCH**
		SGED-GJRGARCH
SGED-APARCH		
JSU-GARCH	**JSU-GARCH**	**JSU-GARCH**
JSU-GJRGARCH	**JSU-GJRGARCH**	**JSU-GJRGARCH**
JSU-APARCH		
JSU-FGARCH	**JSU-FGARCH**	
SGT-GARCH		
SGT-GJRGARCH	**SGT-GJRGARCH**	**SGT-GJRGARCH**
SGT-APARCH		

Panel 2

Probability distributions	
SKST	**5**
JSU	**8**
SGT	**3**

Panel 3

Volatility specifications	
GARCH	**6**
GJRGARCH	**6**
FGARCH	**4**

Table 6: The first panel shows the preferred models under each of the three Dominance criteria. The second panel shows the number of appearances of each probability distribution in the preferred models under at least two Dominance criteria. The third panel shows the number of appearances of each volatility specification in the preferred models under at least two Dominance criteria.

	D1 Rank	D2 Rank	D3 Rank	Median rank	Average rank
N-GARCH	28	28	28	**28**	**28**
N-GJRGARCH	27	27	27	**27**	**27**
N-APARCH	26	26	26	**26**	**26**
N-FGARCH	25	25	25	**25**	**25**
ST-GARCH	18	17	18	**18**	**17.7**
ST-GJRGARCH	20	20	20	**20**	**20**
ST-APARCH	24	23	24	**23**	**23.7**
ST-FGARCH	23	24	23	**24**	**23.3**
SKST-GARCH	2	2	2	**2**	**2**
SKST-GJRGARCH	8	9	8	**8**	**8.3**
SKST-APARCH	13	13	13	**12**	**13**
SKST-FGARCH	6	16	6	**6**	**9.3**
SGED-GARCH	10	10	10	**10**	**10**
SGED-GJRGARCH	5	11	5	**11**	**7**
SGED-APARCH	15	14	15	**14**	**14.7**
SGED-FGARCH	19	19	19	**19**	**19**
JSU-GARCH	1	1	1	**1**	**1**
JSU-GJRGARCH	4	3	4	**3**	**3.7**
JSU-APARCH	12	7	12	**7**	**10.3**
JSU-FGARCH	7	4	7	**4**	**6**
SGT-GARCH	9	6	9	**9**	**8**
SGT-GJRGARCH	3	5	3	**5**	**3.7**
SGT-APARCH	11	8	11	**8**	**10**
SGT-FGARCH	17	12	17	**17**	**15.3**
GHST-GARCH	14	15	14	**15**	**14.3**
GHST-GJRGARCH	22	21	22	**21**	**21.7**
GHST-APARCH	21	22	21	**22**	**21.3**
GHST-FGARCH	16	18	16	**16**	**16.7**

Table 7a: Model ranks according to each Dominance criterion, as well as the median and the mean of those ranks. In case of ties, the same rank is assigned to each model.

Probability distributions	
N	**26.5**
ST	**21.3**
SKST	**7**
SGED	**13.5**
JSU	**3.8**
SGT	**9.8**
GHST	**18.5**
Volatility specifications	
GARCH	**11.9**
GJR-GARCH	**13.6**
APARCH	**16**
FGARCH	**15.9**

Table 7b: Average of the median ranks.

II.3.5. Dominance by asset class

STOCK MARKET INDICES

D1	D2	D3
SKST-GJRGARCH		
		SKST-FGARCH
SGED-GJRGARCH	**SGED-GJRGARCH**	**SGED-GJRGARCH**
	SGED-APARCH	
		JSU-GARCH
JSU-GJRGARCH	**JSU-GJRGARCH**	**JSU-GJRGARCH**
	JSU-APARCH	
	JSU-FGARCH	**JSU-FGARCH**
SGT-GJRGARCH	**SGT-GJRGARCH**	**SGT-GJRGARCH**
	SGT-APARCH	
GHST-GJRGARCH		
GHST-APARCH		
GHST-FGARCH		

Table 8a: Models selected according to each Dominance criteria for stock market indices.

INDIVIDUAL STOCKS

D1	D2	D3
ST-GARCH		
ST-GJRGARCH		
SKST-GARCH	**SKST-GARCH**	**SKST-GARCH**
SKST-GJRGARCH		
SGED-GARCH		
JSU-GARCH	**JSU-GARCH**	**JSU-GARCH**
JSU-GJRGARCH		
JSU-APARCH	**JSU-APARCH**	
SGT-GARCH	**SGT-GARCH**	**SGT-GARCH**
SGT-APARCH		
SGT-FGARCH		
GHST-GARCH	**GHST-GARCH**	**GHST-GARCH**
GHST-GJRGARCH		
GHST-FGARCH	**GHST-FGARCH**	

Table 8b: Models selected according to each Dominance criteria for individual stocks.

INTEREST RATES

D1	D2	D3
	JSU-GARCH	**JSU-GARCH**
	JSU-FGARCH	
	GHST-GARCH	**GHST-GARCH**
		GHST-GJRGARCH
	GHST-APARCH	
GHST-FGARCH	**GHST-FGARCH**	**GHST-FGARCH**

Table 8c: Models selected according to each Dominance criteria for interest rates.

COMMODITIES

D1	D2	D3
	ST-GJRGARCH	ST-GJRGARCH
ST-APARCH		
	ST-FGARCH	ST-FGARCH
		SKST-GARCH
SKST-GJRGARCH		
SKST-APARCH		
	SKST-FGARCH	SKST-FGARCH
SGED-GJRGARCH		SGED-GJRGARCH
SGED-APARCH		
SGED-FGARCH		
JSU-GJRGARCH		
JSU-APARCH		
SGT-GJRGARCH		
SGT-APARCH		
	SGT-FGARCH	SGT-FGARCH

Table 8d: Models selected according to each Dominance criteria for commodities.

EXCHANGE RATES

D1	D2	D3
		N-FGARCH
ST-GARCH	ST-GARCH	
SKST-GARCH	SKST-GARCH	SKST-GARCH
SKST-FGARCH		
SGED-GARCH		
JSU-GARCH	JSU-GARCH	JSU-GARCH
JSU-GJRGARCH	JSU-GJRGARCH	JSU-GJRGARCH
	JSU-FGARCH	
SGT-GARCH		
SGT-GJRGARCH		SGT-GJRGARCH

Table 8e: Models selected according to each Dominance criteria for exchage rates.

SUMMARY

	Indices	Stocks	Interest rates	Exchange rates	Commodities	*Total*
N-GARCH						0
N-GJRGARCH						0
N-APARCH						0
N-FGARCH					D3	1
ST-GARCH		D1			D2,D3	3
ST-GJRGARCH		D1		D2,D3		3
ST-APARCH				D1		1
ST-FGARCH				D2,D3		2
SKST-GARCH		D1,D2,D3		D3	D1,D2,D3	7
SKST-GJRGARCH	D1	D1		D1		3
SKST-APARCH				D1		1
SKST-FGARCH	D3			D2,D3	D1	4
SGED-GARCH		D1			D1	2
SGED-GJRGARCH	D1,D2,D3			D1,D3		5
SGED-APARCH	D2			D1		2
SGED-FGARCH				D1		1
JSU-GARCH	D3	D1,D2,D3	D2,D3		D1,D2,D3	9
JSU-GJRGARCH	D1,D2,D3	D1		D1	D1,D2,D3	8
JSU-APARCH	D2	D1,D2		D1		4
JSU-FGARCH	D2,D3		D2		D2	4
SGT-GARCH		D1,D2,D3			D1	4
SGT-GJRGARCH	D1,D2,D3			D1	D2,D3	6
SGT-APARCH	D2	D1		D1		3
SGT-FGARCH		D1		D2,D3		3
GHST-GARCH		D1,D2,D3	D2,D3			5
GHST-GJRGARCH	D1	D1	D3			3
GHST-APARCH	D1		D2			2
GHST-FGARCH	D1	D1,D2	D1,D2,D3			6

Table 9a: Number of times that each model has been chosen as a preferred VaR model for a given asset class. The table displays the Dominance criteria under which the model was selected.

Probability distributions	
N	**1**
ST	**9**
SKST	**15**
SGED	**10**
JSU	**25**
SGT	**16**
GHST	**16**
Volatility specifications	
GARCH	**30**
GJRGARCH	**27**
APARCH	**13**
FGARCH	**14**

Table 9b: Summary of models selected for the five asset classes by probability distribution and volatility specification.

III. FIGURES

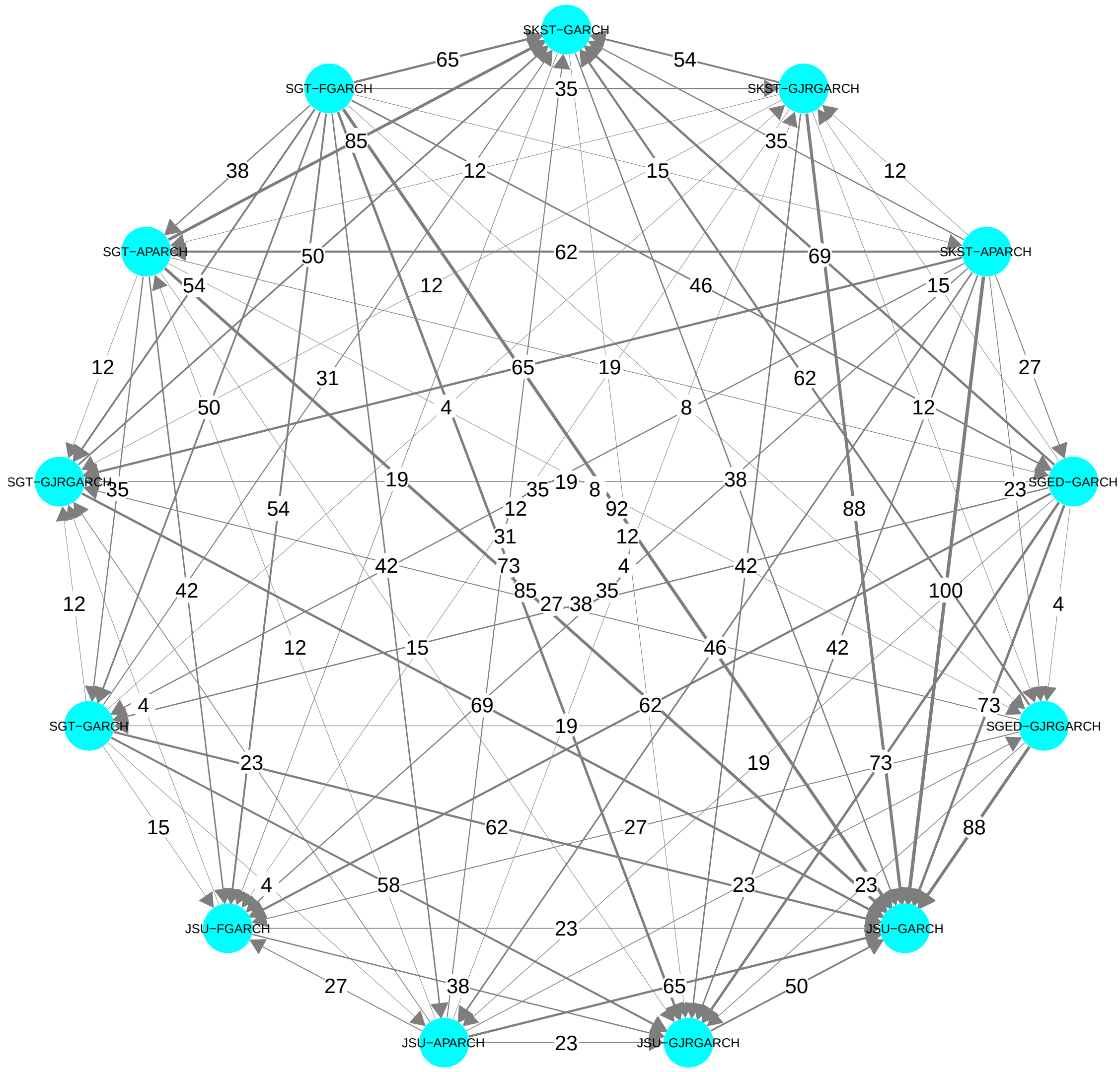

Figure 1: Dominance relationship among models based on aggregate results across the four VaR validation tests, according to Definition 2. Each arrow head points to a model that dominates the model where the arrow originates. The numbers on the arrows show the difference between the number of tests in which the p-value in the dominant model is higher than in the dominated model and the number of tests in which the opposite happens, i.e. the $I2_{12}$ indicator introduced in Definition 2. The values of $I2_{12}$ are shown as a percentage of the highest value of the indicator obtained between any two models.

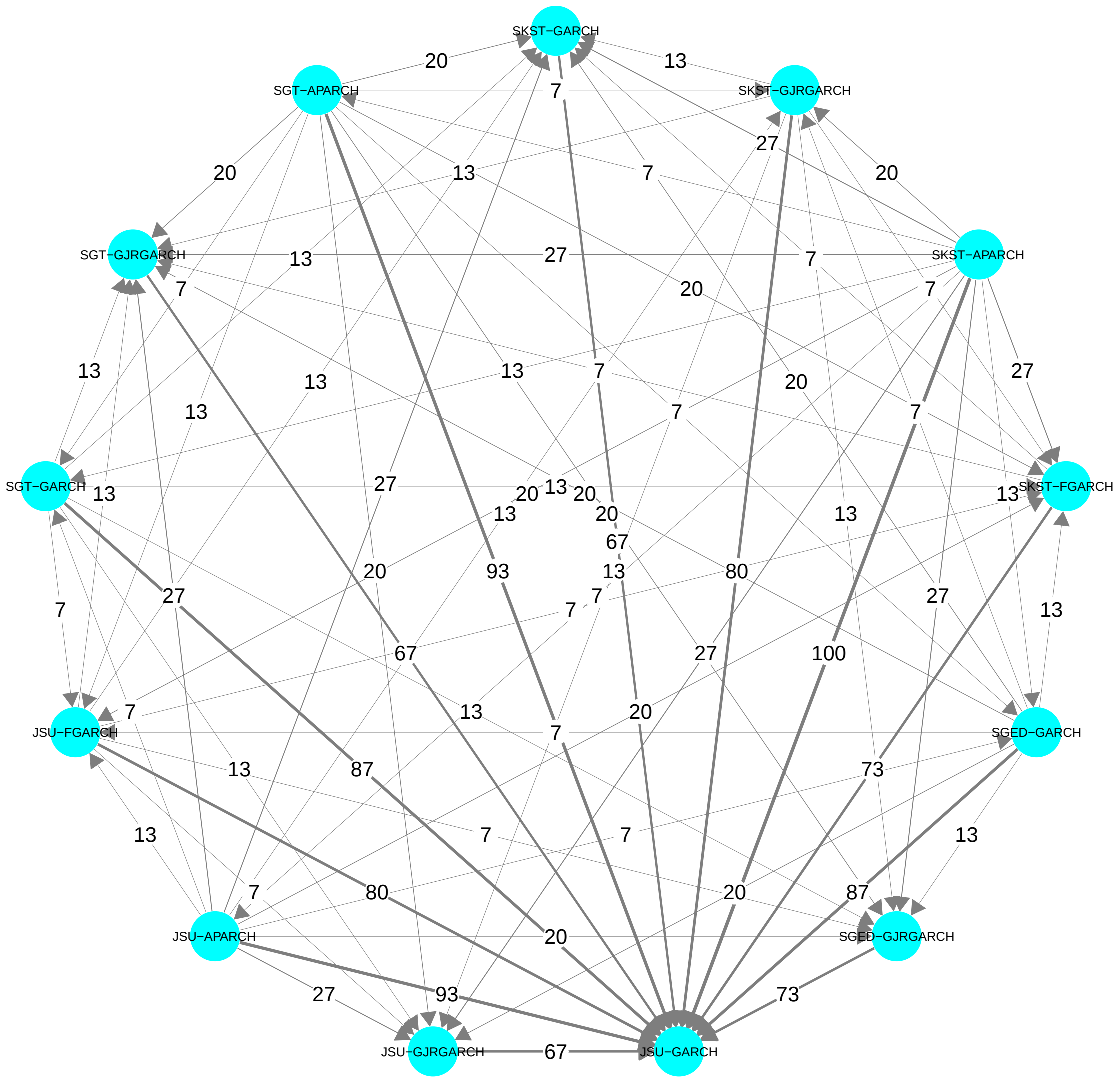

Figure 2: Dominance relationship among models based on aggregate results across the four VaR validation tests, according to Definition 3. Each arrow head points to a model that dominates the model where the arrow originates. The numbers on the arrows show the intensity of dominance measured by the indicator $I3_{12}$, which compares the sum of differences in p-values when $pv1_i < pv2_i$ with the sum of differences in p-values when $pv2_i < pv1_i$. Dominance requires a positive value for that difference. The values of the $I3_{12}$ indicator are shown as a percentage of the largest difference obtained between any two models.